ChinaXiv合作期刊 图书情报工作
第63卷 第11期 2019年6月

数字人文背景下图书馆人文数据组织与重构*

欧阳剑[1] 彭松林[2] 李臻[2]
[1] 广西民族大学文学院 南宁 530006 [2] 广西图书馆 南宁 530024

摘要：［目的/意义］数据是数字人文研究的基础和核心之一，图书馆人文数据组织与重构不但能提高数字资源的利用率，而且能拓展图书馆人文数据服务，可极大地促进数字人文科学的发展，也是图书馆知识型专业服务的具体体现，有利于提供更高层次领域的服务。［方法/过程］分析数字人文研究中的人文数据特点及人文学者研究对人文数据的需求，认为图书馆需从人文数据的完整性、可计算性、可用性及重用性、可发现以及获得性等角度出发进行人文数据组织与重构。［结果/结论］要克服人文数据碎片化带来的零散、不系统的弊病，必须采用数据复原与重构的方式恢复或重建人文数据所蕴含的知识之间的联系，采用数据化、数据融合、数据关联及发布等手段，最终实现知识单元的精细粒度化、知识组织的语义化、知识呈现的可视化。

关键词：数字图书馆 数字人文 人文数据 数据组织 知识重构
分类号：G203
DOI：10.13266/j.issn.0252-3116.2019.11.002

1 引言

数字人文给传统的人文社会学科研究提供了新的研究方法和研究范式[1]，从根本上改变人文学者的研究方式，让研究对象能以数据的形式呈现出来，而且能利用数字技术对数据对象进行分析处理[2]，对人文学者的学科管理和分享他们的研究产生了意义深远的影响[3]，研究人员越来越多地要求大规模地访问具有版权或许可的数据，以进行各种形式的计算研究（如文本挖掘、数据挖掘、机器学习）。在数字人文研究过程中，图书馆作为资源存储库，经过多年的发展，在人文学科领域的数据资源建设已具相当规模，图书馆建立了庞大的人文学科领域数字资源，为数字人文研究提供了一系列信息基础设施，数据与数据服务已成为图书馆服务的延伸[4]，在去中介化的趋势下，图书馆急需从数字馆藏到数字数据、从数据管理到数据服务、从数据呈现到数据分析的转变，面向数字人文研究的服务对图书馆来说既是挑战更是机遇，成为图书馆转型的契机，也将成为图书馆服务新的增长点。如何将这些数据组织重构为适合数字人文研究所需的人文数据是图书馆展开人文数据服务的前提与关键，而人文数据组织与重构是图书馆提供数字人文服务的重要基础。

2 数字人文背景下图书馆人文数据服务面临的机遇与挑战

信息环境的更新迭代不仅仅是理念上的跃进，更是现实需求的凸显。图书馆支持人文学者开展人文计算研究，提供资源与服务已取得了良好的效果，图书馆从最初参与数字化项目的合作，发展到现在与研究人员和供应商协商文本挖掘权及数据成果发布和托管，并为数字人文研究提供研究空间与场所，图书馆在数字人文研究中发挥了至关重要的作用，多年来，图书馆一直是数字人文研究重要的合作者[5]，但更多是将图书馆作为信息资源与信息服务的提供者[6]，数字人文研究致力于促进信息资源的广泛获取和共享、科研数据的处理和研究方法创新、促进学术沟通、加强学习和教学，同时提升文化信息资源的公众影响力，而这些恰恰也是图书馆工作和发展的意义所在，目标的一致性决定了图书馆是数字人文的天然合作伙伴[7]。图书馆作为知识、信息和数据的存储库，已经在数据存储、组织、文本挖掘和元数据标准等数据管理、服务方面积累

* 本文系国家社会科学基金项目“图书馆古籍文献的数字人文开发与应用模式研究”（项目编号：17XTQ003）和广西社会科学基金项目“面向数字人文研究的大规模古籍语料可视化分析与挖掘”（项目编号：15DTQ001）研究成果之一。
作者简介：欧阳剑（ORCID：0000-0001-5867-2852），研究馆员，博士，硕士生导师，E-mail：oyjjj@163.com；彭松林（ORCID：0000-0001-6304-1094），副馆长，讲师，硕士；李臻（ORCID：0000-0003-3981-7782），公共数字文化建设中心副主任，副研究馆员，硕士。
收稿日期：2018-09-21 修回日期：2018-12-17 本文起止页码：15-24 本文责任编辑：王传清

了丰富的经验[8]，数字人文研究为其成功介入到跨学科数据的管理活动中，与人文领域、计算机领域的学术团体建立密切的合作伙伴关系提供了一个独特的机遇[9]。

数字人文由人文计算发展而来，数字人文最显著的特点就是借助计算机进行量化分析，数据是数字人文研究的基础和核心之一，数字人文领域的研究使数据驱动研究成为主流[10]，数字人文研究要求人文数据具有集成化、细粒化、关联化及可计算化。随着数字图书馆的发展，大量的图书、报纸、期刊、照片、绘本、乐曲、古籍、图像以及视频等人文资料被数字化，形成数量庞大、种类繁多、具有较高价值的数字化资源，数字化文档资料、数据库和检索系统等逐渐成为人文研究的基础平台，图书馆数字资源是人文学科研究的沃土，虽然图书馆已在哲学、历史学、文学、语言学、艺术学、人类学等人文社科领域有丰富的数字资源，但图书馆人文数据分散、孤立、封闭而难以被利用的现实一直制约着图书馆在数字人文研究中的作用。图书馆所储存的文本、图像、音频和深度标引及描述它们的元数据通常是数字人文学者的研究对象，但数字化的信息资源，并未真正改变使用者利用文献的方式，数字化文献无法从“读”转变为“分析”，因此，目前图书馆参与数字人文研究的活动有限，作为数字人文主要数据管理者和提供者，图书馆有必要为人文学者提供必要的人文数据，将人文研究学者从繁杂的资料收集、整理和辨伪工作中解脱出来。

图书馆开展面向数字人文研究服务势在必行，是学科馆员服务及嵌入式服务理念的延伸，是知识型专业服务的具体体现，更是图书馆转型创新趋势，将成为图书馆服务新的增长点。多年来，图书馆界也比较重视人文数据组织与重构，建设了庞大的人文数据库，早在1990年，美国国会图书馆的美国记忆等标志性项目开始探索文本、动态图像和音频的大规模数字化[11]，HathiTrust也一直致力于信息资源的保存与共享，并提供对数百万文本作品的访问，通过数据胶囊的形式提供人文数据服务[12]，我国的CADAL数字图书馆也开放了近250万图书。尽管提供了丰富的人文数据库，但图书馆长期以来仅限于展示其人文数据（书籍、图像等），离适合数字人文研究的人文数据还有一定差距。图书馆所存储的庞大数据资源是数字人文发展的重要基础，如何将这些数据组织、重构为适合数字人文研究所需的人文数据是图书馆展开数据服务的前提与关键，这也是图书馆开展数字人文研究服务的基础。图书馆数据组织与重构不但能提高数字资源的利用率，而且能拓展图书馆高层次服务领域，而高质量的数据又能保障数字人文研究的快速展开，提高研究的效率与质量。

人文数据的研究也越来越引起学者的关注，对人文数据的组织与重构进行了比较广泛的探讨，C. Schöch对数字人文研究中的“数据”不同类型进行了详细的分析[13]，T. Padilla则对人文数据收集的完整性、表现形式及访问等进行了研究[14]；人文数据监护也引起了学者的注意[15-16]，一些会议（沙龙）也对构造面向可计算的人文数据及人文数据的重用问题进行了广泛的探讨[17]，而人文资料的数据化已经进行了大量的实践，形成了一定规模的人文数据库[18-19]。人文数据是数据人文研究的基础之一，人文数据的组织与重构是图书馆提供数据服务的重要任务，也是数字人文背景下的图书馆人文数据服务面临的机遇与挑战。

3 数字人文研究中的人文数据特点与人文学者需求

人文数据主要由计算机处理的可计算化的数字形式编码，主要由格式化数据、文本、图像、音频和视频等组成。图书馆人文数据组织服务与应用研究应以需求为导向，根据人文数据的特点深入调研科研人员的数据需求，遵循数字人文研究应用中的数据获取、标注、比较、取样、阐释与表现方式，以图书馆现有数据资源为基础进行抽取、融合、重组形成若干人文学科研究所需数据，积极创造有利于科研人员沟通和创造的基础人文数据平台。

3.1 数字人文研究中的人文数据特点

在传统人文学科研究的过程中学者的大部分时间耗费在相关材料收集及整理方面，而且人文学科的研究缺乏研究团队，大多以个体研究为主，人文学科研究所需资料是通过长时间积累而成，加上人文学者数字化技术的欠缺使得数据建设的过程周期长，人文数据的个体化色彩通常很强，不同人文学者对同一份资料的解读往往千差万别，难以达成共识；另一方面，记载又往往在质量、体量题材、记录方式、详略程度上极不均匀，导致随之而来的数据经常有大量残缺[20]，因此，从宏观层次来说这就决定了人文数据杂乱且碎片化特点，呈现出非结构化、混乱和隐含、形式各异。

从微观角度来看人文数据具有两个维度，第一个维度描述数据的结构，清晰和显式，第二个维度描述了数据的大小和变化程度[12]。大部分人文数据是形式

化的表格，具有数据结构清晰、属性值大小和变化程度明确的特性，但人文学科中的人文数据也有其特殊性，书本或手稿中的文字或构成绘画的视觉元素虽然是数据，但他们是模拟的非离散数据，难以通过计算分析或转换，语言、文本、绘画和音乐则具有超出物理上可测量的维度的符号系统，这些维度的分析依赖于语义和语用，即在语境中的意义依赖于语境的解释和人工理解标注，因此其可能具有多义性，人文数据增加了研究人员与其研究对象之间关系的复杂性。人文数据的可计算、可量化是数字人文研究的另一大特点，数字人文核心目标是将现代信息技术融入人文领域，从而改变知识的获取、标注、比较、取样、阐释与表现方式，通过设计、计算分析、可视化等手段重塑和改造人文知识，为学者提供更多差异化、规律性、宏观性、趋势性研究的可能和线索，从而扩展学术疆域和潜力。

3.2 人文学者的人文数据需求

数字人文研究对人文数据提出了独特的要求，人文数据的组织与构建很大程度上由学科规范和方法论所决定，人文数据的组织通常需要有人文素养的介入，即需要了解人文数据特点及符合人文学者研究的需求才能确保人文数据的有效性。集成与融合的数字化资料与数据是数字人文研究的基础，人文学者的数字人文研究模式从以“读”文献的方式为主转变为“分析”文献为主，将文献中的描述内容转变为可分析的数据，以此作为人文学科研究的辅助手段，即基于数据的研究[21]。

空间与时间是人类赖以生存和发展的双重维度，也是历代哲人思考和探讨的焦点论题[22]。人文学科研究的对象大多与时间紧密结合，需以时间为主线分析研究对象的演变、形成及发展过程，对历时性的内容变迁深入理解，从空间角度对研究对象从地理空间进行分析和解读，从时空角度分析空间位置的分布组合与变迁，事物关系分析是时间分析和空间分析的再综合，强调事物之间的关系或结构在时间和空间上的固定联系和相互影响，数字人文的研究视角也主要聚焦在研究对象空间、时间及之间的关系上，使得数字人文的研究具有多视角的特性，因此，人文数据需具有多维性，能从时间、空间与对象之间的关系角度描述对象特性。

人文学科研究数据的多维性集成首先是需要将同类研究目的的数据融合，通过对同类研究的结果进行综合分析，以获取新的概念，从而使认识水平提高到一个新的高度，其次是将不同类别、不同目的的研究数据融合，经过对比与数理统计分析，力求反映出各研究主题与其他要素之间的关系，并解释出隐含在其背后的规律。人文数据不但有描述对象的元数据，还有依赖于语境的解释和人工理解的标注数据，更有意义表达的语义数据，既有单一属性的数据，也有事物整体描述的数据，通过将不同层次、不同角度的碎片化数据建立关联，从而形成一个统一的知识表述与构建，重构发现原来内在的知识，或产生新的理解，将碎片化知识整合后，有利于形成系统而完整的知识体系。知识重构是数字人文的重要应用，可进一步激活并再生人类知识。

数据驱动的研究范式被越来越多地应用在人文学科中，将不同来源、海量的、各种不同类型的、结构化和非结构化、特点性质的数据在逻辑上或物理上有机地集中和展现，提供人文数据共享与重用。人文数据的组织与重构可认为是数字人文研究的基础，早在 1949 年，B. Roberto 使用电脑处理神学家 A. Thomas 的全集，半自动地生成出作品中拉丁文字词的索引，其实就是人文数据组织与重构。

4 面向数字人文研究的图书馆人文数据组织及重构基本要素

信息科学家 L. Floridi 将数据定义为最基本的单元，只有当数据具有一些可识别的结构并具有某种意义，它们才能被视为信息[23]，数据可以用许多不同的形式来表示，数据的特殊之处在于它是离散的而不是连续的，人文学科中的数据可认为是对给定对象的某些方面含义有选择地通过机器所能理解及可读的数字来表示与描述[13]。人文数据的加工、组织和解释由学科规范和方法论所决定，人文数据的可加工性使得在其宏观层面上也能够通过微观的数据来测量、识别，使得数字人文学者能将人文数据广泛应用于可视化和数据挖掘[24]。数字人文研究中的人文数据具有的特点及人文学者研究对人文数据的需求构成了图书馆人文数据组织与重构的基本要素，其中主要有人文数据的完整性、可计算性、可用性及重用性、可发现及获得性等。

4.1 人文数据完整性

人文学科研究带有很明显的实证性，研究材料的真实可靠、合理选取材料范围成为人文学者的学术传统[25]，人文研究者极其考证资料的真实性问题及材料的溯源，数字人文背景下的图书馆人文数据组织与重构首先需要做到人文数据的完整性[14]。人文数据完整性主要有两层含义：一是指人文数据的收集、加工、

转换及发布的生命周期内实现人文数据的可溯源;二是指某一类人文数据所收录的覆盖程度。

人文学科研究极为重视材料的真实可靠性,保护人文数据收集、加工、转换及发布的完整性和关键材料的可追溯性,以使人文数据具有批判性的可寻址性,使研究人员能够理解为什么包含和排除某些数据,为什么进行某些转换,谁进行了这些转换,同时使研究人员能够访问用于实现这些转换的代码和工具,就像网络档案一样必须有效地传达这些细微差别,这一关键可寻址性概念在整个研究过程中至关重要,研究人员希望根据人文学科研究的需要来选择、评估和追溯原始材料。人文数据的完整性主要体现在来源、处理和演示 3 个方面,来源是人文数据的出处,做到原始材料的可追溯性;处理是指研究人员对人文数据加工、转换等过程,处理过程中需要保证数据一致性;演示是指用于呈现已处理数据的方法和工具,人文数据收集来源、处理和演示方法直接影响人文学者对数据的可信度。“中国历代人物传记资料库(China Biographical Database, CBDB)”及“中华文明之时空基础架构(Chinese Civilization in Time and Space,CCTS)”等人文数据建设过程中都保留有数据来源记录信息。

多视角、多维度研究早已嵌入到人文学科研究中,多维度分析要求数据能覆盖不同研究视角,将不同来源、各种不同类型的、结构化和非结构化、特点性质的数据在逻辑上或物理上有机地关联,能够辅助人文学者从多层面、多角度来揭示问题。人文学者历来重视个案研究,个案研究属于微观研究,人文研究对个案研究来说,最基本的价值是材料完整性,即某一类个案人文数据所收录覆盖尽量全。随着数字化环境的发展,新生产的可用信息资源越来越多,数字化的资源也越来越多,为人文数据建设提供了充裕的来源,人文数据的完整性为图书馆人文数据的收集、开发及实践提供了明确方向。

4.2 人文数据可计算性

数字人文是人文计算的延续和发展,人文数据可计算性的量化分析是数字人文的核心,也是数字人文研究区别于传统人文学科研究的显著特点。计量分析是对所研究的对象存在的特征进行量化分析,计量分析从对某些具有数字特征的事件作单一变量的统计描述,到多个不具备数字特征的事物或事件进行定量研究[26],参与计算的对象须有明确的计量属性,这就要求人文数据及知识颗粒化、属性的独立性,从多角度进行精细化元数据加工与标注,以揭示文献形式和内容的多种属性,分解出一系列量值便于人文计算,便于从空间与时间角度再现其横纵细节特征,人文数据建设过程中的“数据化”任务其实就是文献内容和形式的多种属性描述与标注。数字人文研究的对象是可计算的基础数字化对象,数字人文研究分析的不仅仅限于描述及标注性的元数据分析、数字数据分析或形式化数据的量化分析,语言及历史文本挖掘、考古及文化遗产图像分析、舞蹈视频捕捉及运动分析、网络社交数据分析也是分析的对象,因此,人文数据应包括人文研究中所有能进行人文分析的对象属性与特征。

4.3 人文数据可用性及重用性

人文数据是数字人文研究的基础之一,数字人文研究虽然具有跨学科特性,但同时也是属于专业性极强的研究,人文数据往往具有极强的专用性,使得人文数据的应用场景具有较大的局限性,而实现人文数据的通用性与适用性是图书馆追求的目标,为了使已建成的人文数据应用于更多研究场景,人文数据可用性及重用性对数字人文来说非常重要,人文数据对象本身采取何种形式存储、发布直接影响到人文数据的可用性及重用性。

人文数据通常以一定样式实例化,一组通用的格式和数据结构可以更好地支持数字人文的研究和教学,随着人文数据文档各种标准化的建立,极大地促进了人文学科的研究和教学,文本编码倡议(text encoding initiative,TEI)为电子形式的文本材料定义一系列的通用标准,TEI 被世界各国以文本为基础的人文研究广泛使用[27],虽然 TEI 被视为更高级用户的核心格式,但 TEI 底层的一系列 XML 文件中保存的数据通常不太适用于数字人文,因此,近年来人文数据越来越迎合人文学者的实际需要,开始制定收集转换策略以便于数字人文中的各种数据之间的转换[28],在人文数据组织与构建时,需要明确在功能层面决定哪些人文数据表格最受人文研究者欢迎,按相对常见的数字人文研究工具和方法存在共同的数据格式要求进行转换,将需要的人文数据集转换以便更好地支持想要计算与集合交互的用户,能根据人文研究工具和方法中的数据格式要求生成更容易使用的数据格式,如 Access、Excel 等格式的数据,提高了人文数据的易用性。

目前,人文数据大多是科研项目所产生的,随着人文项目的结束,研究团队解散,数据的维护和更新就成为主要问题,人文数据的维护也将成为数据重用的一个重要影响因素,因此,人文数据的长期存储与监护也成为人文数据的可用性及重用性关键,数据监护(data

curation)服务的兴起为数字人文研究的数据的长期保存、交换与被更广泛地获取和重用提供了保障[29]。

4.4 可发现及获得性

人文数据集差异巨大,众多数据混杂使得数据孤岛的现象依然存在,人文数据的建设的主要目的是服务于人文学者,提高人文数据的可发现性、可访问性、可获得性等对图书馆数据服务来说至关重要。

为了支持人文数据的可访问及获得性人文数据建设、组织、管理过程中的数据揭示是基础,主要体现在人文数据的著录、索引、本体及语义等内在描述,人文数据揭示与描述侧重于人文数据特征的表示,描述了数据的特征和之间的各种关联。而人文数据浏览、导航及相关检索等外在数据发现工具同样不可或缺,它们直接面向用户,是连结人文研究者与人文数据之间的桥梁与纽带,正如 Google 新推出数据集搜索工具 Dataset Search (https://toolbox. google. com/dataset-search),其发布凸显了 Google 对数据检索的重视,借助于工具对人文数据进行检索,增强了人文数据的可发现性。人文数据重现和透明度的价值日益受到重视,人文数据组织、开发的重点是实现人文数据的访问,人文研究者最终目的是需要获得数据,当前的访问方式差异巨大,从静态层次结构的简单网页、XML 文件、XSL 文件等传统数据的传统获取途径,到利用 Github 及 FTP 进行文本集合访问,发展到越来越多的应用程序编程接口(API)等获取途径,简化了人文数据的获取方式。

5 面向数字人文研究的图书馆人文数据组织及重构模式与方法

5.1 面向数字人文研究的图书馆人文数据组织与重构模式

在数字图书馆观念中,图书馆组织的对象主要是数字化的信息资源,信息资源是按照学科分类体系以高度结构化的方式存在,以资源为中心进行组织,信息与信息之间存在一个严格的层次结构,构成静态的“金字塔式”信息结构模式(见图1),这样的信息资源对于人文学者来说还是属于文献层次的信息,数字化文献无法实现从“读”转变为“分析”的作用,严格的层次结构难以满足人文学者的学科研究的需求,与数字人文研究所要求的人文数据的完整性、可计算性、可用性及重用性、可发现以及获得性等具有巨大差异。数字人文研究中的人文学者通常专注于专题性的研究,需要分析覆盖某个专题的全部人文数据,因而必须考虑个性化的人文数据定制,要求人文数据被组织成容于访问且可用于计算,同时又要满足人文研究中的多角度比较、取样、阐释等,往往横跨多个数据集,要求人文数据具有很强的关联性,因此,相关人文数据的大规模融合以及资源的细粒度、关联性重建,成为图书馆支撑人文研究的建设重点[30]。

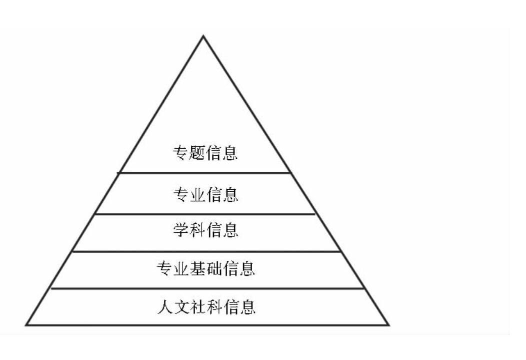


图1 “金字塔式”信息结构

从概念上来说,数据与信息处于不同层次,只有那些数据具有一些可识别的结构并具有某种意义才能被视为信息,传统“金字塔式”信息结构模式使人文数据处于碎片化状态,数据碎片化的本质就是数据之间的联系被网络或人为地切断了,某个数据的知识点与相关知识点处于分离的状态,被人为地孤立。而数字人文研究以研究任务为中心,人文数据对于研究者来说应屏蔽各数据结构与存储,人文数据服务于所研究的课题,能方便、快捷地获取到所需要的相关人文数据,人文数据之间以研究课题为中心形成“蛛网式”数据结构(见图2),数据之间经过组织与重构变成能反映特定学科或领域研究的“智慧数据”,使数据之间建立起关联满足人文学者的多维分析视角需要。

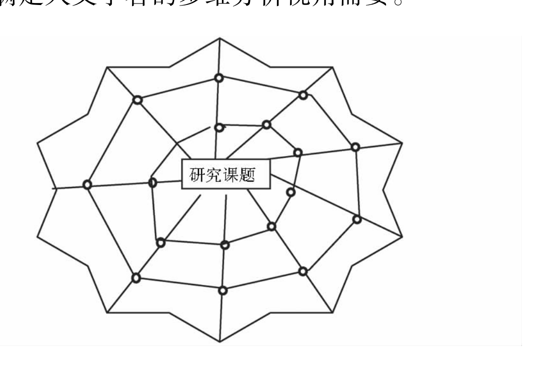
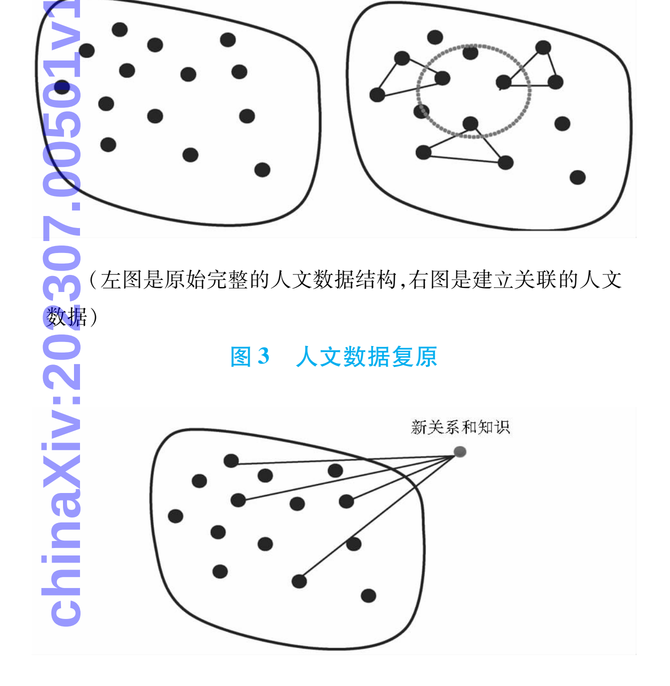


图2 “蛛网式”数据结构

图书馆要克服人文数据碎片化带来的零散、不系统的弊病,必须恢复与重建人文数据所蕴含的知识之间的联系。人文数据的组织与重构主要有人文数据复原与人文数据重构这两种形式。人文数据复原是指按照原有的学科知识体系重建原来的系统化数据与知识

结构，侧重于人文数据的数据化与数据之间关联的建立，首先完成文本、图像、音视频的颗粒化深度标引与描述，形成原始完整的人文数据结构，在此基础上按照他们之间的关系建立起关联的人文数据（见图 3）；人文数据重构是指不必严格参照原有的学科知识体系，而是按照人文学者研究的需要，以个性化研究课题的人文数据结构进行组织与重建，在原始人文数据中发现原来没有的关系和知识，重构更有利于解决人文学者面临的真实问题与场景还原，更有利于知识创新（见图 4）。比较典型的“威尼斯时间机器”（Venice Time Machine）项目通过数字化的古地图、专著、手稿和乐谱等大量文件中离散的知识与数据重构了威尼斯千年历史[31]。



（左图是原始完整的人文数据结构，右图是建立关联的人文数据）

图 3　人文数据复原

（在原始人文数据中发现原来没有的关系和知识）

图 4　人文数据重构

5.2　面向数字人文研究的图书馆人文数据组织与重构方法

人文数据组织与重构有异于传统的数字图书馆数字资源整合，数字资源整合也称为数字资源集成，是在各种数字资源自主性、分布性、异构性的基础上，运用各种集成技术和手段将各类数字资源集成在统一的利用环境下，实现“一步到位”的检索，面向数字人文研究的图书馆人文数据组织与重构方法就不同于数字资源整合，人文数据融合与图书馆资源整合有本质的区别，人文数据组织与重构则是经过分析、综合、转换以及发布等一系列人文数据加工处理工作而构建完整、权威的人文数据集，并建立起人文数据之间的关联，不仅仅包含数字化，更包含文本、图像、音视频的多角度的颗粒化深度标引与元数据描述、数据化、数据融合、知识关联等工作，最终实现知识单元的细粒度化、知识组织的语义化、知识呈现的可视化。

5.2.1　数据化　数字人文研究活动包含数字化、数据化、数据管理、数据计算分析等，数字化作为数字图书的最终产物与形态，距离数字人文的可计算性还有一定距离，因此有必要将电子形态进一步转换为可识别的文本与可分析的数据，以便做进一步的计量，因此，数据化是数字人文研究的一项基础工作，数据化的核心工作是重组文献内容，置入使用者所建立的新的文本或数据结构中，即文献的结构化[19]，数字人文研究中的收集、标注直接以数据为管理对象，而数字人文研究中的比较、取样、计算分析等则依赖于数据分析与解读。数据化包括光学字符识别文本过程，使文献资源便于文本分析与挖掘等，这是数据化的一种初始阶段，数据化也包括重组文献内容，即将文献内容转化为可制表分析的量化形式[32]，转换为可量化分析的数据。

目前，大部分文献数据化工作处于光学字符识别文本阶段，手稿及古籍自动识别依然面临重大的技术挑战，随着数字人文的发展，更多的文献内容重组及形式化的工作也相继展开，如哈佛大学费正清中国研究中心、北京大学人文社会科学研究院、台湾“中央”研究院历史语言研究所合作开发的中国历代人物传记资料库，哈佛大学的地理分析中心和复旦大学的历史地理研究所合作的中国历史地理信息系统、上海交通大学的“中国地方历史文献数据库”以及台湾大学数位人文研究中心的多个数据库等相继出现，以文献内容为基础，从数字人文的理念出发将人文学者所需要的文献内容转化为可制表分析的量化形式呈现，实现知识多角度的精细化、关联化，满足数字人文的完整性、可计算性、可用性及重用性、可发现及获得性等，这是数据化的未来发展方向。

5.2.2　数据融合　传统的量化分析通常是对单一数据源进行深入的追踪和分析，分析人员对数据的来源和结构有一定的控制和深层的了解。数字人文研究则特别强调人文数据的重用性与多视角的取样分析，形成有效的多视角分析数据集是数字人文研究必须面对的一个瓶颈，也是大数据背景下人文学科研究的基础。数据化只是实现了传统数字人文素材向数字世界的映射，能够被计算机所存储、处理和展示，人文数据的多维性则要求通过信息及知识单元的方式来组织，从而

能够构造一个模拟领域应用的环境,在这个过程中,数据融合成为不可或缺的一步。数据融合使人文学者可以轻松驾驭多样、多源的数据,便于进行多维度挖掘和分析,帮助人文学者发现新规律、新价值。经过多年的发展,数字图书馆资源平台存储了大量可计算的基础数据,作为数字人文的重要数据来源,因此,数据的复用和重组是极为重要的,数据重用和重组是人文数据组织与重构的主要任务,融合不同图书馆的不同人文数据对数字人文研究致关重要。

人文数据融合是对同一研究对象相关的多个属性数据采用一定的模式与方法,生成一个新的、更能有效表示该研究对象的综合数据集或获得新的隐性知识,将单一数据或不同类别的多源数据加以综合,消除多源信息之间可能存在的冗余和矛盾,加以互补,改善研究对象信息提取的及时性和可靠性,提高数据的使用效率。人文数据融合首先连接所需多源数据库并获取相关数据,研究和理解所获得的数据,在梳理和清理数据的基础上进行数据转换和建立结构,实现数据组合、数据整合和数据聚合并建立分析数据集。

从融合形式来说,人文数据融合主要有异构融合、多源融合、多模融合 3 种形式,异构融合是指将结构数据、半结构数据和非结构等不同存储结构形式的人文数据进行融合;多源融合是针对来自于不同的学科领域和数据源的人文数据进行融合;多模融合是指对文本、影像、语音等不同的人文数据形式进行融合。数据融合胜于数据仓库和数据一体化在于它能包容多源数据,可将不同属性并与某一潜在的对象存在一定隐含关联的多源数据集融合形成一个新的数据集,甚至新知识,融合提高了人文数据的互补性和完整性。如图 5 所示:

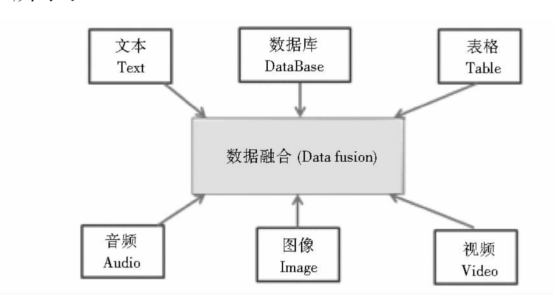


图 5　人文数据融合形式

从融合层次来说,多源人文数据融合可分为:数据层融合、特征层融合及决策层融合(见图 6)[33],数据层融合是指直接对采集及加工的原始数据进行简单组合,数据层融合是数据融合的最简单方法,也是数据融合的基础。特征层融合则是在数据层融合的基础上进行特征提取,然后对特征信息进行综合分析和处理,特征层融合是基于特征拼接的方法。决策层融合则通过关联处理进行决策层融合判决,最终获得联合推断结果,决策层融合是基于语义信息融合的方法,包括了多视角、基于概率学模型的方法、基于相似度的方法、以及迁移学习的方法。

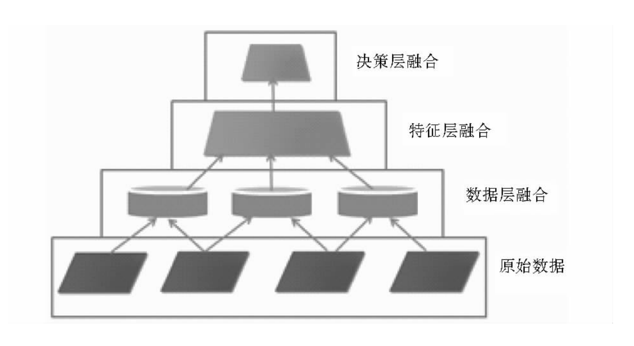


图 6　人文数据融合层次

5.2.3　数据关联及发布　人文数据的可访问性和可用性问题是数字人文研究的重要挑战之一,人文学者往往需要将研究查询从一个人文数据集转换到另一个人文数据集,甚至允许直接交叉数据集查询,然而人文学科研究成果的多样性,以及人文科学研究工作通常围绕单个个体或小团体的努力进行组织,使得人文学科的数据访问、共享与重用似乎是一个无法实现的目标。目前,关联数据技术最有可能填补这一空白,改善人文数字的访问及重用的局限性[34]。人文数据集是各种人文数据通过建立关联而成的数据集,数据集从古代地图到书目记录,到绘画、音视频,到古文字分析,再到插图事实等对象,它们之间存在各种紧密关系,其中一些相互关联,因此需要聚合、集成并提供交叉搜索,查找实体和作品之间的链接,构建叙述,分析数据[35]。语义技术和关联数据使大规模数字人文的协作和聚合研究成为可能,其中关联数据和知识图谱增强了机器可读、机器可理解性,通过关联数据技术构建关系明确的语义本体,实现基于文献知识内容的揭示,增强了人文数据重用和与外部数据的互联,形成了互连和分散的全局知识网络,实现了一种人文数据链接的新模式,使人文数据作为一个整体被人文学者所利用。近年来,上海图书馆应用关联数据技术对家谱数据及历史地理数据的开放应用进行大量实践[36],采用关联数据使得从以图书馆为中心的知识组织系统向跨领域公开可用和易于访问的知识图转变,提高了人文数据的可用性和重用性。

人文数据领域特定知识结构的开发和管理是人文

学科的重要元素,在图书馆和数字人文领域内,知识的概念图深深植根于知识组织系统,主要涉及异构源的知识(潜在的)语义索引、分类或查询、导航以及可视化等,知识组织系统在图书馆领域具有数百年的传统,在元数据描述中它们被用于组织资源并促进资源 发现和检索。随着关联数据的出现,知识组织系统经历了数字化转型并进入了互联网领域,2012 年,随着“Google Knowledge Graph”的出现[37],知识图谱立刻受到学界及工业界的普遍关注,并成为研究的热点,给图书馆领域的知识组织带来了新的变革。知识图谱是近年来知识组织领域的研究热点,是一种以语义网络为基础的新型海量知识管理和服务模式,知识图谱旨在描述客观世界的概念、实体、事件及其间的语义关系[38]。构建知识图谱的主要目的是获取大量的、让计算机可读的知识以及实体及其相关属性 - 值对,实体之间通过关系相互联结,构成网状的知识结构,增强知识单元之间的关联,实现用户主题检索需求,从而真正实现语义检索[39]。

知识图谱技术渊源已久,很长一段时间以来,学术界一直关注如特定领域内的地方或人等权威数据的管理,这些知识图谱一直遵循与图书馆知识组织方案类似的模式:它们越来越多地根据关联数据原则发布,并与网络上的其它知识图谱相关联,使用共享的语义概念[40]。知识图谱本身就是一种语义化的表示方式,具有明显的语义网特征。基于数字人文研究的知识图谱构建适应了数字人文研究的需要,通过知识图谱可以对这些信息资源进行语义标注和链接,对同一研究对象的多个属性数据采用知识图谱的形式,以需求为导向在统一系统平台中对数字化文献所蕴涵的多重信息进行多角度的揭示和重组,语义联系组合成纵横交错的多维结构,建立以知识为中心的资源语义集成服务,突破传统的应用模式,充分展示人文知识的最大价值。对同一研究对象的多个属性数据采用知识图谱的形式,可生成一个新的、更能有效表示该研究对象的综合数据集或获得新的隐性知识。跨域知识图谱正如图书馆和数字人文领域中出现的那样,开辟了一个全新的研究机会。

近年来,知识图谱在国内外的数字人文项目中越来越受到重视,展现出了巨大的应用前景。笔者近年来致力于古籍知识图谱的构建,通过与古籍知识密切相关的古籍编撰者、籍贯、时间(年代)、编撰方式、藏书机构等要素,围绕从时间(年代)、空间、关系等角度进行多维关系构建,通过整理中国、日本及欧美主要国家的近 200 万种古籍数据,形成了中国古籍知识图谱[41],方便了古籍知识(潜在的)语义索引、分类或查询、计算分析及可视化等,从数字人文研究应用的维度来说,从强大知识关联性方面有助于考察版本源流,理清流变脉络,还能够通过古籍知识图谱分析责任者、编撰时间、编撰方式、版本特征等多种维度的相关性分析,从而进一步揭示古籍数据背后隐藏的丰富文化、历史等知识,突破了传统的以古籍单一数据源统计分析的模式,通过规则推理技术可以获取数据中存在的隐含知识,通过古籍责任者空间信息可视化分析功能,为文学地理的空间环境分析提供了新的研究方式,提升了古籍文献目录知识服务的价值。

6 结语

数据是数字人文研究的基础和核心之一,在数字人文研究过程中,图书馆作为资源存储库,庞大的数字资源成为人文研究重要的人文数据来源,在去中介化的趋势下,图书馆急需从数字馆藏到数字数据、从数据管理到数据服务、从数据呈现到数据分析的转变,面向数字人文研究的服务对图书馆来说既是机遇更是挑战,成为图书馆转型的契机,而面向数字人文研究的图书馆人文数据组织与重构则成为关键,图书馆需要了解数字人文研究的人文数据特点及文学者研究对人文数据的需求,从人文数据的完整性、可计算性、可用性及重用性、可发现以及获得性等角度出发进行人文数据组织与重构,图书馆要克服人文数据碎片化带来的零散、不系统的弊病,必须恢复或重建人文数据所蕴含的知识之间的联系,利用数据化、数据融合、数据关联及发布等手段,最终实现知识单元的细粒度化、知识组织的语义化、知识呈现的可视化。图书馆人文数据组织与重构不但能提高数字资源的利用率,而且能拓展图书馆人文数据服务,可极大地促进数字人文科学的发展,也是图书馆知识型专业服务的具体体现,有利于提供更高层次领域的服务。

参考文献:

[1] 王晓光. 数字人文:概念、现状与思考[EB/OL]. [2018 - 06 - 26]. http://meeting. lib. szu. edu. cn/conference/zh -hans/information? v =07000003.

[2] LEONARD P. Mining large datasets for the humanities [EB/OL]. [2018 - 08 - 26]. http://library. ifla. org/930/1/119 -leonard -en. pdf.

[3] 赖德伯格 - 科克斯. 挑战数字图书馆和数字人文科学[M]. 朱常红,译. 广西:广西师范大学出版社,2010.

[4] 秦健. 数据与数据服务:图书馆据服务的延伸 [EB/OL]. [2018 - 06 - 26]. http://society. library. sh. cn/.

[5] Special report: digital humanities in libraries[EB/OL].[2018－07－09]. https://americanlibrariesmagazine.org/2016/01/04/special-report-digital-humanities-libraries/.

[6] SCHAFFNER J, ERWAY R. Does every research Library need a digital humanities center? [EB/OL]. [2018－07－27]. http://oclc.org/research/publications/library/2014/oclcresearch-digital-humanitiescenter-2014-overview.html.

[7] SPIRO L. Why digital humanities [EB/OL]. [2018－07－01]. http://digitalscholarship.files.wordpress.com/2011/10/dhglca-5.pdf.

[8] ACRL. Top trends in academic libraries[EB/OL]. [2018－07－01]. http://crln.acrl.org/content/75/6/294.full#xref-ref-68-1.

[9] VANDEGRIFT M, VARNER S. Volving in common: creating mutually supportive relationships between libraries and the digital humanities [J]. Journal of library administration, 2013(53): 67－78.

[10] Data driven: digital humanities in the library [EB/OL].[2018－07－18].http://dhinthelibrary.wordpress.com/.

[11] American memory: remaining collections[EB/OL].[2018－12－15].http://memory.loc.gov/.

[12] HathiTrust research center data capsule v1.0: an overview of functionality [EB/OL].[2018－09－18]. https://scholarworks.iu.edu/dspace/handle/2022/18936.

[13] Big? smart? cean? messy? data in the humanities[EB/OL].[2018－07－09].http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/.

[14] PADILLA T. Humanities data in the library: integrity, form, access [EB/OL]. [2018－07－09]. http://dlib.org/dlib/march16/padilla/03padilla.print.html.

[15] Humanities data: a hands-on approach [EB/OL].[2018－12－13]. http://digital.humanities.ox.ac.uk/dhoxss/2016/workshops/dhcuration.

[16] FLANDERS J, MUNOZ T. An introduction to humanities data curation[EB/OL]. [2018－12－13]. http://guide.dhcuration.org/contents/intro/.

[17] Shaping humanities data: use, reuse, and paths toward computationally amenable cultural heritage collections [EB/OL].[2018－12－13].https://dh2017.adho.org/abstracts/670/670.pdf.

[18] 徐力恒. 唐代人物资料的数据化:中国历代人物传记资料库(CBDB)近年工作管窥[M]// 包伟民,刘后滨.《唐宋历史评论》第三辑.北京:社会科学文献出版社,2017:20－32.

[19] 赵思渊.地方历史文献的数字化、数据化与文本挖掘:以《中国地方历史文献数据库》为例[J].清史研究,2016(4):26－35.

[20] 徐力恒.中国历史人物大数据[J].中国计算机学会通讯,2018,14(4):19－24.

[21] 刘炜,叶鹰.数字人文的技术体系与理论结构探讨[J].中国图书馆学报,2017,43(5):32－41.

[22] 梅新林.文学地理学:基于"空间"之维的理论建构[J].浙江社会科学,2015(3):122－136,160.

[23] FLORIDI L. Information: a very short introduction[M]. Oxford: Oxford University Press, 2010:22－25.

[24] PADILLA T G, HIGGINS D. Library collections as humanities data: the facet effect[J]. Public services quarterly,2014,10 (4): 324－35.

[25] 李醒东,刘晓红.科学研究中资料的真实性诉求背后——论人文研究的价值追求[J].当代教育科学,2007(11):43－44.

[26] 人文科学的定量分析试探[J].文艺理论研究,1992(1):43.

[27] SPERBERG-MCQUEEN M. Text encoding and enrichment[M]// The humanities computing yearbook 1989－90. Oxford: Oxford University Press, 1991.

[28] SCHREIBMAN S, SIEMENS R, UNSSWORTH J. A companion to digital humanities[M]. Oxford: Blackwell, 2004.

[29] POOLE A H. A greatly unexplored area: digital curation and innovation in digital humanities[J]. Journal of the Association for Information Science & Technology, 2017, 68(7):1－10.

[30] 李欣,张毅,汪志莉.图书馆异构特藏资源整合的数字人文研究需求[J].数字图书馆论坛,2017(11):48－53.

[31] ABBOTT A. The 'time machine' reconstructing ancient Venice's social networks[J]. Nature, 2017, 546(7658):341－344.

[32] 舍恩伯格.大数据时代:生活、工作与思维的大变革[M].周涛,译.杭州:浙江人民出版社,2013: 104.

[33] 欧阳剑.面向数字人文研究的多源数据融合[EB/OL].[2018－12－13]. http://society.library.sh.cn/adls2016.

[34] HOEKSTRA R, MEROÑO-PEÑUELA A, DENTLER K, et al. An ecosystem for linked humanities data[EB/OL].[2018－12－13]. https://www.semanticscholar.org/paper/An-ecosystem-for-Linked-Humanities-Data-Hoekstra-Mero%C3%B1o-Pe%C3%B1uela/e82d876d5e7ef8a09430d38ee340da86acff0d0c? tab = abstract.

[35] Linked open data for cultural heritage and digital humanities [EB/OL].[2018－07－09].https://ontotext.com/linked-open-data-cultural-heritage/.

[36] 夏翠娟,刘炜,陈涛,等.家谱关联数据服务平台的开发实践[J].中国图书馆学报,2016,42(3):27－38.

[37] SINGHAL A. Introducing the knowledge graph: things, not strings [EB/OL].[2018－07－21]. https://goo.gl/U168iz.

[38] 李涓子,侯磊. 知识图谱研究综述[J]. 山西大学学报(自然科学版),2017(3):454－459.

[39] 刘峤,李杨,段宏,等.知识图谱构建技术综述[J].计算机研究与发展,2016,53(3):582－600.

[40] HASLHOFER B, ISAAC A, SIMON R. Knowledge graphs in the libraries and digital humanities domain[EB/OL].[2018－12－13]. https://arxiv.org/pdf/1803.03198.pdf.

[41] 欧阳剑.基于数字人文研究的大规模古籍文献知识图谱构建[R].北京:中国图书馆学会,2017.

作者贡献说明:

欧阳剑:论文主题选取、结构构思与撰写论文;

彭松林:修改论文;

李臻:修改论文。

Organization and Reconstruction of Library's Humanities Data Under the Background of Digital Humanities

Ou Yangjian[1] Peng Songlin[2] Li Zhen[2]
[1] Guangxi University for Nationalities, Nanning 530006
[2] Guangxi Library, Nanning 530024

Abstract: [**Purpose/significance**] Data is one of the foundation and core of digital humanities research. The organization and reconstruction of humanities data in library can not only improve the utilization of digital resources, but also expand the library humanities data services, which can greatly promote the development of digital humanities science. It is also a concrete manifestation of library knowledge-based professional services, which is conducive to providing services in higher level areas. [**Method/process**] This paper analyzes the characteristics of humanities data in digital humanities research and the demand for humanities data by humanities researchers, and considers that the library should be based on the integrity, calculation, usability and reusability of humanities data to organize and reconstruct humanistic data of library. [**Result/conclusion**] To overcome the fragmented and unsystematic ills caused by the fragmentation of humanities data, it is necessary to use the link between the knowledge of restoring or rebuilding the humanities data, using the means of data, data fusion, data association and publication, and finally realizing the fine granularity of knowledge unit, the semantic of knowledge organization and the visualization of knowledge presentation.

Keywords: digital library digital humanities humanities data data organization knowledge reconstruction

iConference 2019 在美国马里兰大学帕克分校举行

2019 年 3 月 31 日 -4 月 3 日,2019 年(第十四届)iConference 会议在美国马里兰大学帕克分校举行,本次会议由马里兰大学帕克分校主办,雪城大学和马里兰大学巴尔的摩郡分校协办,会议获得美国国家科学基金会、美国计算研究协会、爱墨瑞得出版社、爱思唯尔、MDPI 出版社、摩根和克莱普尔出版社、NVIVO 公司、匹兹堡大学计算与信息学院、肯塔基大学信息科学学院、台湾大学图书资讯学系、田纳西大学诺克斯维尔分校信息科学学院等机构赞助。本次会议的主题是启智、包容、启迪(inform,include,inspire),旨在探讨 21 世纪启智的意义,如何扩大信息革命的覆盖面,并思考如何更好地启迪个人和组织在这个快速变化的知识社会中利用信息。全球 60 余位 iSchool 学院院长,近 600 名学者参加此次会议,中国人民大学、武汉大学、南京大学、北京大学、中山大学、南开大学、南京理工大学、河北大学、西北师范大学、云南师范大学、华中师范大学等中国高校师生参加了本次会议。本次会议共接收 77 篇论文和 91 篇海报,接收的论文由施普林格计算机科学讲义(Springer's Lecture Notes in Computer Science)收录,并在伊利诺伊大学图书馆的 IDEALS(Illinois Digital Environment for Access to Learning and Scholarship)开放获取。本次会议上,共有 3 名专家作大会主题报告,分别是密歇根大学信息学院 W. K. Kellogg 社区信息教授 Kentaro Toyama 博士作题为"技术的扩增法则及其对 iSchool 的意义(Technology's Law of Amplification, and What It Means for iSchools)"报告,互联网档案馆创始人 Brewster Kahle 博士作题为"开放图书馆:百万在线图书的受控数字借阅"(Openingour Libraries: Millions of Books Online through Controlled Digital Lending)报告,第 14 任美国国会图书馆馆长 Carla Diane Hayden 博士作题为"数字时代的图书馆:怎么办?"(Libraries in the Digital Age: Now What?)报告。本次会议评选出"Understanding the Role of Privacy and Trust in Intelligent Personal Assistant Adoption"获得 LEE DIRKS 最佳论文奖,"Characterizing Same Work Relationships in Large-Scale Digital Libraries"获最佳短文奖,"Algorithmic Accountability in Surveillance Regulation"获最佳海报奖,"Disrupting the Coming Robot Stampedes: Designing Resilient Information Ecologies"、"Troubled Worlds: Bringing Bodies and the Environment into Computing Research, Practice, and Pedagogy"、"Human Security Informatics: A Human-centered Approach to Tackling Information and Recordkeeping Issues Integral to Societal Grand Challenges"分获蓝天论文(BLUE SKY PAPER)一、二、三等奖。

(华中师范大学曹高辉供稿)